A note on the choice and the estimation of Kriging models for the analysis of deterministic computer experiments

David Ginsbourger, Delphine Dupuy, Anca Badea, Laurent Carraro, Olivier Roustant

A note on the choice and the estimation of Kriging models for the analysis of deterministic computer experiments

David Ginsbourger *, Delphine Dupuy*, Anca Badea*, Laurent Carraro*, Olivier Roustant*

April 3, 2008

Abstract

Our goal in the present work is to give an insight on some important questions to be asked when choosing a Kriging model for the analysis of numerical experiments. We are especially concerned about the cases where the size of the design of experiments is small relatively to the algebraic dimension of the inputs. We first fix the notations and recall some basic properties of Kriging. Then we expose two experimental studies on subjects that are often skipped in the field of computer simulation analysis: the lack of reliability of likelihood maximization with few data, and the consequences of a trend misspecification. We finally propose an example from a porous media application, with the introduction of an original Kriging method in which a non-linear additive model is used as external trend.

Keywords: Metamodeling, Kriging, Maximum Likelihood, Deterministic Drift, Additive Models

1 Linear predictors for spatial interpolation of numerical simulators

We study a deterministic numerical simulator as a function $z : D \subset \mathbb{R}^d \to \mathbb{R}$, where $\mathbf{x} \in D$ is the vector of inputs variables. We denote the set of the design

*Département 3MI , Ecole Nationale Supérieure des Mines, 158 cours Fauriel, 42023 Saint-Etienne (France), tel. +33 04 77 49 97 57, e-mail: ginsbourger@emse.fr

points (or "design") by $\mathbf{X} = \{\mathbf{x^1}, ..., \mathbf{x^n}\}$ and by $\mathbf{Z} = \{z(\mathbf{x^1}), ..., z(\mathbf{x^n})\}$ the set of simulator responses associated with $\mathbf{X}$. Kriging is a class of methods coming from the field of geostatistics [15, 3], known as *linear optimal prediction* in classical statistics. It provides at each point $\mathbf{x} \in D$ a prediction $\hat{Z}(\mathbf{x})$ linearly depending on $\mathbf{Z}$, where the weights depend on the design and on the Kriging model but not on the observations. The way the weights are defined varies as a function of the type of Kriging -Simple (SK), Ordinary (OK), Universal (UK), etc- and many parameters such as the trend functions, the covariance kernel and their own parameters: threshold (or "sill"), scales, nugget, etc... denoted by the r-dimensional vector $\boldsymbol{\psi}$. In the following, we will concentrate on the parameters of sill and scale ($r = 2$), denoted respectively either by $\psi_1,\ \psi_2$ or by $\sigma^2,\ p \in [0, +\infty[$. Most classic Kriging types (including SK, OK, UK, and more) can be interpreted as random process interpolation relying on the assumption that:

$$\forall \mathbf{x} \in D,\ z(\mathbf{x}) = t(\mathbf{x}) + \varepsilon(\mathbf{x}) \tag{1}$$

where t is a numerical deterministic function and $\varepsilon(\mathbf{x})$ is one path of a centered stationary Gaussian Process (GP) with known stationary covariance kernel $k : h \in \mathbb{R}^d \longrightarrow k(h) \in \mathbb{R}$. t is generally known up to a set of parameters or a semi-parametric structure to be estimated within Kriging. Several founder works [19, 7] on the application of Kriging to computer simulations start off with an extremely simplified version of (eq.1). They assume that the trend is an unknown constant (Ordinary Kriging, i.e. $t(x) = \mu \in \mathbb{R}$) and that k is a generalized exponential kernel [20], letting the stochastic part of (eq.1) account for the variability of z. Then the covariance parameters $\boldsymbol{\psi}$ are estimated by maximizing the Gaussian likelihood of the observations $\mathbf{Z}$. On the other hand, recent approaches [8, 14] try to take advantage of more complex trends, from linear and polynomial functions to Fourier series. In other respects, [13] as well as [16] present an application of bayesian analysis to Kriging interpolation of computer codes.

The motivation of this article is to raise some basic questions that should become crucial when applying Kriging techniques with few observations regarding the dimension of inputs, which is quite often the case in numerical simulation. The two coming sections, based on toy experiments, put a focus on the estimation of the covariance parameters $\boldsymbol{\psi}$ and on the choice of the trend t. The two following sections are dedicated at presenting an original combination of additive models and Simple Kriging, with a heuristic fitting methodology. The efficiency of this technique is illustrated on a 3-dimensional example from a porous media simulation test case.

2 Fitting covariance parameters by MLE with a small sample

The Maximum Likelihood (ML) estimation method is widely used in Kriging to choose covariance parameters on the basis of observations. Following the assumptions from (eq. 1), ML estimation relies on the maximization of the density of the observed values $\mathbf{Z}$, seen as a function of the vector $\boldsymbol{\psi}$:

$$L(\boldsymbol{\psi}; \mathbf{Z}) := f(\mathbf{Z}|\boldsymbol{\psi}) = (2\pi)^{-\frac{n}{2}} \det(K_{\boldsymbol{\psi}})^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{Z}-\mathbf{t})^T K_{\boldsymbol{\psi}}^{-1}(\mathbf{Z}-\mathbf{t})} \tag{2}$$

where $K_{\boldsymbol{\psi}}$ is the covariance matrix of $Z(\mathbf{X}) = \{Z(\mathbf{x^1}), ..., Z(\mathbf{x^n})\}$ provided that $\boldsymbol{\psi}$ is the true vector of covariance parameters, and $\mathbf{t}$ is the vector of values of t at $\mathbf{X}$. The obtained result $\widehat{\boldsymbol{\psi}} = argmax_{\boldsymbol{\psi}}\{L(\boldsymbol{\psi}; \mathbf{Z})\}$ is closely depending on $\mathbf{Z}$, i.e. on the observed realization of $Z(\mathbf{X})$. The behaviour of $\widehat{\boldsymbol{\psi}}$ relatively to $\boldsymbol{\psi}$ when the sample of observations fluctuates is a of importance. We recall that $\mathbf{Z}$ is assumed to be one realization of a multivariate Gaussian random vector with given trend, covariance structure, and covariance parameters $\boldsymbol{\psi}$. Then $L(.; \mathbf{Z})$ becomes a random function (fig. 1), and $\widehat{\boldsymbol{\psi}} = argmax_{\boldsymbol{\psi}}\{L(\boldsymbol{\psi}; \mathbf{Z})\}$ becomes a random vector as well.

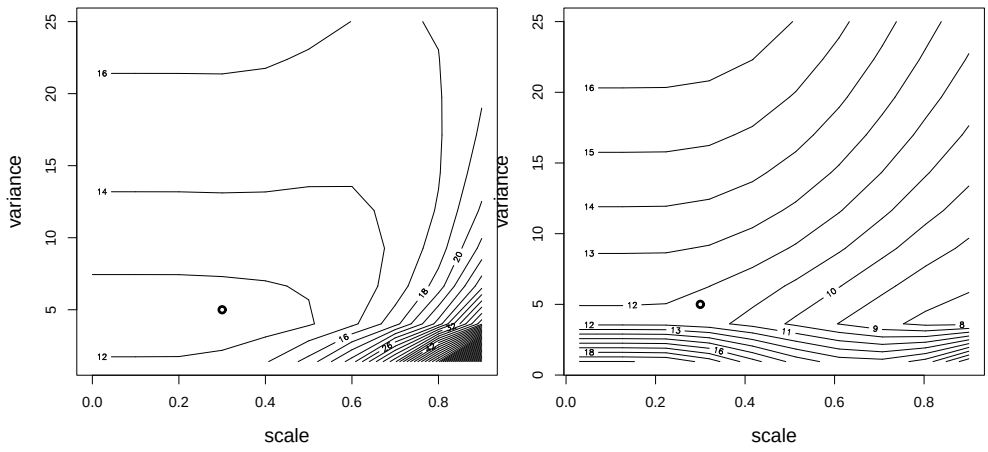

Figure 1: Two realizations of the random function $-2 \ln L(.; \mathbf{Z})$ corresponding to two simulated response values $\mathbf{Z}$, both with a Gaussian covariance kernel (c_g defined hereafter) and covariance parameters $\boldsymbol{\psi} = (5,\ 0.3)$. **Left:** ML estimates are close to the actual parameters (bold dot) : $\widehat{\boldsymbol{\psi}} \approx \boldsymbol{\psi}$. **Right:** ML fails to locate the actual parameters: $\widehat{\boldsymbol{\psi}} \neq \boldsymbol{\psi}$.

The distribution of $\widehat{\boldsymbol{\psi}}$ has been studied in detail within the theory of likelihood [23, 12]. A first order Taylor expansion leads to an asymptotic result [2] based on Fisher's Information Matrix $\mathcal{I}(\boldsymbol{\psi})$ (denoted by FIM in the sequel):

$$\begin{cases} \widehat{\boldsymbol{\psi}} \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{\psi}, \mathcal{I}(\boldsymbol{\psi})^{-1}) \\ \mathcal{I}(\boldsymbol{\psi}) = \left(\mathbb{E}\left[\frac{\partial \ln(L(.;\mathbf{Z}))}{\partial \psi_i}(\boldsymbol{\psi}) \frac{\partial \ln(L(.;\mathbf{Z}))}{\partial \psi_j}(\boldsymbol{\psi}) \right] \right)_{i,j \in [1,r]} \end{cases} \tag{3}$$

In many computer experiments, one first picks a covariance kernel from a parametric family: Gaussian, Exponential, Matèrn, etc...(the Gaussian covariance kernel is often chosen for its simplicity and regularity properties) and the associated covariance parameters are then automatically fitted by ML. However, the efficiency and robustness of this estimation method when few data are available are rarely discussed. Our concern is to check in what measure the first order asymptotic results hold with small samples. To do so, we computed the theoretical FIM of $\mathbf{Z}$:

$$\forall i,j \in [1,r] \ (\mathcal{I}(\boldsymbol{\psi}))_{ij} = \frac{1}{2} tr\left(K_{\boldsymbol{\psi}}^{-1} \frac{\partial K_{(.)}}{\partial \psi_i}(\boldsymbol{\psi}) K_{\boldsymbol{\psi}}^{-1} \frac{\partial K_{(.)}}{\partial \psi_j}(\boldsymbol{\psi}) \right) \tag{4}$$

To obtain comparable results for different values of ψ, we introduce a relative inverse FIM: $(\mathcal{J}(\boldsymbol{\psi}))_{ij} = (\mathcal{I}^{-1}(\boldsymbol{\psi}))_{ij}/(\psi_i \psi_j)$. $\mathcal{J}$ is in fact the asymptotical covariance matrix of $\frac{\widehat{\boldsymbol{\psi}}}{\boldsymbol{\psi}}$, where the division is made component by component. We conduct experiments with vectors taken from simulated monodimensional Gaussian Processes to compute empirical means and variances of the ML estimators. For each simulation, we compute covariance parameters estimated by ML and the Integrated Squared Error (ISE) between simulated ($z(\mathbf{x})$) and interpolated ($\widehat{Z}(\mathbf{x})$) data:

$$\text{ISE} = \frac{1}{\text{vol}(D)} \int_D |z(\mathbf{x}) - \widehat{Z}(\mathbf{x})|^2 d\mathbf{x} \tag{5}$$

where vol(D) is Lebesgue's measure of the set D. ISE is approximated by averaging the squared errors on a fine grid (i.e. 200 points). We finally collect the averages and variance matrices of the relative values of the estimated covariance parameters, the averages and variances of ISE (ISE is random since it depends on the realization z), and the covariances between ISE and $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$. The latter two indicators are not presented in the tables. We focus here on GPs with covariance kernels $c_g(h) = \sigma^2 e^{-\frac{h^2}{p^2}}$ (Gaussian) and $c_e(h) = \sigma^2 e^{-\frac{|h|}{p}}$ (Exponential). The covariance parameters reduce to $\boldsymbol{\psi} =$

$(\sigma^2, p) \in]0, +\infty[\times]0, +\infty[$ and the design $\mathbf{X}$ is chosen among uniform subdivisions of $[-1, 1]$: $\mathbf{X_n} = \{-1, -1+\frac{2}{n-1}, \ldots, -1+\frac{2(n-2)}{n-1}, 1\}$ $(n \in \mathbb{N}\backslash\{0, 1\})$. We restrict our experiments to the designs $\mathbf{X_5}$ and $\mathbf{X_{10}}$ with both c_g and c_e, and covariance parameters $\psi_1 = \sigma^2 \in \{5, 10\}$, and $\psi_2 = p \in \{0.3, 0.4, 0.5, 0.6\}$.

Table 1: ML and ISE values on 1000 simulated realizations of GPs with Gaussian covariance function, for relative parameters $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$, $i = 1, 2$ and for $\mathbf{X} = \mathbf{X_5}$. The second column shows that the relative ML estimates are almost unbiased even with 5 observations. On the contrary, a comparison between the third and fourth columns illustrates that the ψ_i^{rel} are clearly more dispersed than given by the asymptotical approximation based on the FIM.

$\boldsymbol{\psi}$	$\mathbb{E}\left[\left(\psi_i^{rel}\right)_i\right]$	$Var\left[\left(\psi_i^{rel}\right)_i\right]$	asymptotical $Var\left[\left(\psi_i^{rel}\right)_i\right]$	$\mathbb{E}$[ISE]
$\begin{pmatrix} 5 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.034 \\ 0.018 \end{pmatrix}$	$\begin{pmatrix} 1.105 & 0.277 \\ 0.277 & 1.270 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	4.976
$\begin{pmatrix} 5 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.147 \\ 0.042 \end{pmatrix}$	$\begin{pmatrix} 1.329 & 0.501 \\ 0.501 & 0.976 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	3.287
$\begin{pmatrix} 5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.222 \\ 0.033 \end{pmatrix}$	$\begin{pmatrix} 4.037 & 0.757 \\ 0.757 & 0.679 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	1.947
$\begin{pmatrix} 5 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.187 \\ 0.027 \end{pmatrix}$	$\begin{pmatrix} 2.058 & 0.504 \\ 0.504 & 0.421 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	0.706
$\begin{pmatrix} 10 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.131 \\ 0.006 \end{pmatrix}$	$\begin{pmatrix} 3.334 & 0.867 \\ 0.867 & 1.564 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	10.138
$\begin{pmatrix} 10 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.083 \\ 0.106 \end{pmatrix}$	$\begin{pmatrix} 1.645 & 0.484 \\ 0.484 & 0.862 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	6.398
$\begin{pmatrix} 10 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.166 \\ 0.024 \end{pmatrix}$	$\begin{pmatrix} 1.343 & 0.440 \\ 0.440 & 0.629 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	3.678
$\begin{pmatrix} 10 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.256 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} \mathbf{14.960} & 0.963 \\ 0.963 & 0.392 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	1.459

In the case of a Gaussian covariance, we observe a negative relative bias [1] (-3.4% to -25.6%) in the estimation of $\psi_1 = \sigma^2$. This bias is decreasing with the number of design points $\#\mathbf{X}$ (see table 2 where the negative relative bias varies between -4.2% and -7.5%), which seems in accordance with the asymptotic unbiasedness of MLE. On the other hand, the relative bias of $\hat{\psi}_2$ has a small order of magnitude when $\#\mathbf{X} = 5$ and slightly oscillates around 0 when $\#\mathbf{X} = 10$.

The empirical covariance matrices of the ML estimates offer some surprising results. In particular, the relative variances of $\hat{\psi}_1$ present huge fluctuations: they vary sometimes of an order of more than 10 between two samples of 1000 realizations issued from the same GP; for instance by resimulating a GP with

[1] Mind the fact that by negative relative bias we understood an overestimation of $\boldsymbol{\psi}$.

Table 2: MLE and ISE measures on 1000 simulated GP realizations with Gaussian covariance kernel, for $\mathbf{X} = \mathbf{X_{10}}$. The aproximation based on Fisher's Information Matrix is still underestimating the estimation variances but is less unprecise than in table (1).

$\boldsymbol{\psi}$	$\mathbb{E}\left[\left(\psi_i^{rel}\right)_i\right]$	$Var\left[\left(\psi_i^{rel}\right)_i\right]$	asymptotical $Var\left[\left(\psi_i^{rel}\right)_i\right]$	$\mathbb{E}$[ISE]
$\begin{pmatrix} 5 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.054 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.432 & 0.105 \\ 0.105 & 0.085 \end{pmatrix}$	$\begin{pmatrix} 0.297 & 0.057 \\ 0.057 & 0.033 \end{pmatrix}$	0.177
$\begin{pmatrix} 5 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.042 \\ -0.019 \end{pmatrix}$	$\begin{pmatrix} 0.424 & 0.058 \\ 0.058 & 0.024 \end{pmatrix}$	$\begin{pmatrix} 0.340 & 0.044 \\ 0.044 & 0.014 \end{pmatrix}$	0.009
$\begin{pmatrix} 5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.067 \\ -0.013 \end{pmatrix}$	$\begin{pmatrix} 0.46 & 0.051 \\ 0.051 & 0.013 \end{pmatrix}$	$\begin{pmatrix} 0.362 & 0.036 \\ 0.036 & 0.008 \end{pmatrix}$	0.0004
$\begin{pmatrix} 5 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.075 \\ -0.007 \end{pmatrix}$	$\begin{pmatrix} 0.728 & 0.059 \\ 0.059 & 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.375 & 0.032 \\ 0.032 & 0.005 \end{pmatrix}$	4.e-05

$\psi = (10, 0.4)$ and $\#\mathbf{X} = 5$ we obtain $Var\left[\left(\psi_i^{rel}\right)_i\right] = \begin{pmatrix} 43.555 & 3.242 \\ 3.242 & 0.971 \end{pmatrix}$. Since it is in contradiction with normality and the order of magnitude given by (eq.4), we shall analyze this phenomenon in detail. First, we observe that the extreme values of $Var[\hat{\psi}_1]$ are caused by some outliers, highly perturbating the non-robust estimate of variance. Second, the histogram in (fig.2) illustrates that the distribution of the $\hat{\psi}_1$'s is rather lognormal than normal. Finally, the comparison with the relative FIM shows that the empirical variance of $\hat{\psi}_1$ is clearly larger than predicted by the second order Fisher approximation, in particular with the smallest designs.

Concerning the relative variances of $\hat{\psi}_2$, the results are much more regular: they decrease monotonically with ψ_2 and with $\#\mathbf{X}$, both for the empirical and theoretical quantities. Once again, the empirical variances tend to match the theoretical variances as $\#\mathbf{X}$ grows, even if the first ones are still typically two times larger than the second ones for a sample of size 10. In other respects, both tables illustrate some fundamental properties of the mean squared error. Obviously decreasing with $\#\mathbf{X}$, the ISE is also decreasing with the range ψ_2 and linearly increasing with the variance ψ_1. Finally, we quantify the linear dependence between the underestimation of both covariance parameters by MLE and the ISE (not in the tables). It is worth noticing that ψ_1 and ψ_2 play drastically different roles here: it seems that a bad estimation of ψ_1 is weakly correlated with the ISE. This result seems natural when considering that the OK predictor is not depending on the process variance, see [3]. Conversely, the correlation between the ISE and the relative MLE error on ψ_2 is significantly positive: it varies between 40.1% and 55.7% when $\#\mathbf{X} = 5$ and between 15% and 62.5% when $\#\mathbf{X} = 10$. This coincides with

our previous qualitatives observations of larger ISE when the range is much underestimated.

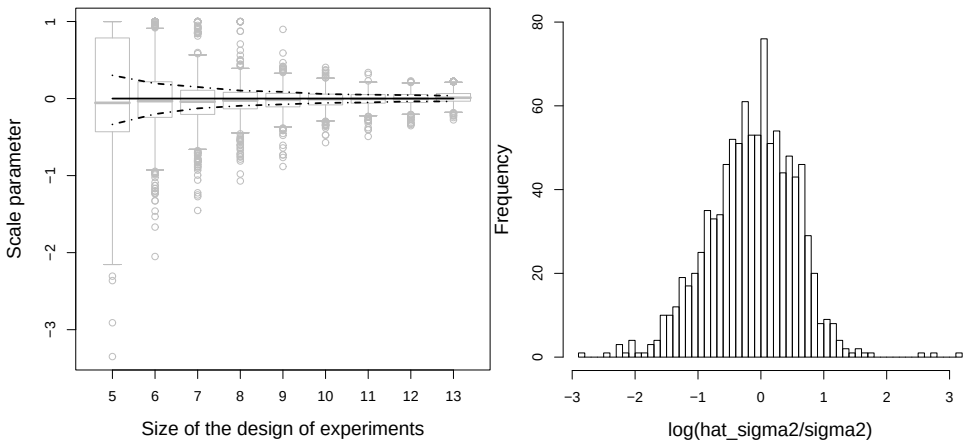

Figure 2: **Left:** Comparison between the experimental law (gray boxplots) and the asymptotic law (black lines) for the scale parameter for increasing size of the design. The boxplots for the experimental laws have been done using 1000 simulations, with Gaussian covariance function of parameters $\boldsymbol{\psi} = (5, 0.5)$. For the asymptotic law the median is represented in continuous line and the first and third quartiles in dashed lines. **Right:** Histogram of the logarithm of relative errors obtained when estimating ψ_1 by ML using 1000 GP realizations. The shape of this histogram suggests that the distribution of relative errors is far closer to a lognormal law than to a Gaussian.

A similar study with exponential covariance function gives very different results both for the bias and the variances of ML estimates (the corresponding tables are not presented here). Indeed, we observe very regular variances of ML estimates while the bias reaches impressive orders of magnitude. However, the behaviour of the ISE and the correlations between ISE and relative MLE errors follow the same sketch as in the Gaussian case.

To sum up this section about ML estimation:

- Fisher's first order asymptotical results must be applied with much care concerning the sample size. More precisely, it has been observed here for $n \leq 5$ that the distribution of the estimated range parameter

is asymmetrical with a higher variance than the inverse of Fisher's information, but quickly stabilizes to a Gaussian when n increases (from 5 to 13).

- On the other hand, the distribution of the estimated variance parameter has a very large right tail but its shape is far from being gaussian when n is very small ($n \leq 5$). Furthermore, these results still hold when n increases (from 5 to 13) and it seems that the Gaussian approximation becomes reasonable only for larger values of n.

Estimating covariance parameters by ML with few data appears to produce very dispersed results. Hence, it seems unreasonable to neglect the uncertainty associated with this phase of estimation when performing Kriging. Bayesian techniques are a way to address this issue [16, 4]. In other respects, [1] investigates an extended Kriging variance taking the estimation of parameters into account; a this stage the latter relies on the first order approximation. To finish with, frequentist approaches based on the maximization of penalized likelihood functions seem very promizing since they provide estimators with the same asymptotic properties as ML in addition to a more robust behaviour with few observations [11].

3 Kriging with trends: a bless or a curse?

Now we wish to examine another difficulty encountered when Kriging based on few data: the selection and the estimation of deterministic trends. In computer experiments, the most commonly used Kriging model seems to be Ordinary Kriging. However, OK reaches one of its limits when the stationarity assumption does not hold any longer, i.e. when non constant trends $t(\mathbf{x})$ are impossible to ignore. In this case, we are back to the general decomposition of (eq.1), where z is assumed to be the sum of a deterministic trend t and one realization of a centered GP ε. At this stage, we may consider several subcases.
If t is known and the parameters of ε have to be estimated, a straightforward solution is to perform Simple Kriging of the residuals $\{z(\mathbf{x}) - t(\mathbf{x})\}_{\mathbf{x}\in D}$.
If t is unknown, it is common to distinguish between a linear and a more general non-linear framework. The case in which t depends linearly on its parameters and ε has a kwown covariance structure has been intensively studied: it is well known as Universal Kriging [14]. When the covariance parameters $\boldsymbol{\psi}$ are known and the trend is a linear combination of some chosen basis functions f_j ($j \in [1, b],\ b \in \mathbb{N}\backslash\{0\}$), the only unknowns are the

parameters of the trend ($\forall j \in [1,b],\ \beta_j \in \mathbb{R}$); indeed, if $t(\mathbf{x}) = \sum_{j=1}^{b} \beta_j f_j(\mathbf{x})$, the β_j's can directly be estimated by Generalized Least Squares (GLS):

$$\hat{\beta}(\psi_2) = (\mathbf{F}^T K_{\boldsymbol{\psi}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T K_{\boldsymbol{\psi}}^{-1} \mathbf{Z} = (\mathbf{F}^T R_{\psi_2}^{-1} \mathbf{F})^{-1} \mathbf{F}^T R_{\psi_2}^{-1} \mathbf{Z} \tag{6}$$

where $\mathbf{F}$ denotes the evaluation of $\mathbf{f}(x) = [f_1(x), \ldots, f_b(x)]$ at the n design points and $R_{\psi_2} = (1/\psi_1) K_{\boldsymbol{\psi}}$ (proportionality since the observations are noise-free) is the correlation matrix of $Z(\mathbf{X})$.
In practice, however, one has seldom the value of the covariance parameters at disposal previous to performing UK. So one has to estimate a model with linear trend and unknown covariance parameters $\boldsymbol{\psi}$ (in the following we will also refer to this case as "UK", like many practitioners do). Hence $\boldsymbol{\psi}$ and β have to be estimated within Kriging. At a first sight, this is likely to create a circularity problem: one needs a known trend to work on the residuals and thus estimate $\boldsymbol{\psi}$. On the other hand, estimating t without taking the residuals into account may lead to unadapted trends (the estimation of the trend parameters would rely on Ordinary Least Squares instead of GLS).

Fortunately, ML estimation gives a way to escape this vicious circle. Assuming, like in section 2 that the covariance parameters to be estimated are $\boldsymbol{\psi} = (\sigma^2, p)$, and using MLE (and the same formula (6) for $\hat{\beta}$), one can get a straightforward formula for $\widehat{\sigma^2}$, explicitly depending on ψ_2:

$$\widehat{\sigma^2}(\psi_2) = (1/n)(\mathbf{Z} - \mathbf{F}\hat{\beta}(\psi_2))^T R_{\psi_2}^{-1} (\mathbf{Z} - \mathbf{F}\hat{\beta}(\psi_2)) \tag{7}$$

By injecting (6) and (7) in the expression of the likelihood, one can obtain a *concentrated likelihood* function $L(\psi_2, \widehat{\sigma^2}(\psi_2), \hat{\beta}(\psi_2))$ which clearly depends only on ψ_2 and which has to be maximized to get $\hat{\psi}_2$. The Kriging predictor with plugged-in covariance parameters is then given by:

$$\hat{Z}_{\hat{\psi}_2}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\beta}(\hat{\psi}_2) + r^T(\mathbf{x}) R_{\hat{\psi}_2}^{-1} \left(\mathbf{Z} - \mathbf{F}\hat{\beta}(\hat{\psi}_2) \right) \tag{8}$$

where $r(\mathbf{x})$ is a vector of correlation values between Z at an unknown point $\mathbf{x}$ and at the points of the design $\mathbf{X}$. Most of the time (apart in Bayesian Kriging) the variability due to the estimation ψ_2 is not propagated, and one uses the regular UK prediction variance.

UK appears as a very convenient means to incorporate known deterministic trends within Kriging. By the way, we will see in the next section that overcoming the circularity problem is not easy in a more general non-linear framework. Now we would like to go one step deeper in practical considerations and raise a naive but complex question which has to be handled in

real-world applications, and particularly in high-dimensional problems: how can one come back to the nature of the trend from raw data? As soon as neither prior information nor obvious graphical clue is available, one has indeed to select a trend on the basis of $(\mathbf{X}, \mathbf{Z})$. What means does he have to do so, and what risk does he run in case of a bad choice? In order to show that these questions are crucial, let us first perform some toy experiments. The set-up is the following. A realization of a one-dimensional GP with known covariance function and parameters is simulated on a regular grid (401 points on $[-1, 1]$) and an affine trend is added; From this set we choose different subsets of points and perform three types of Kriging : OK, UK with linear trend and UK with quadratic trend.

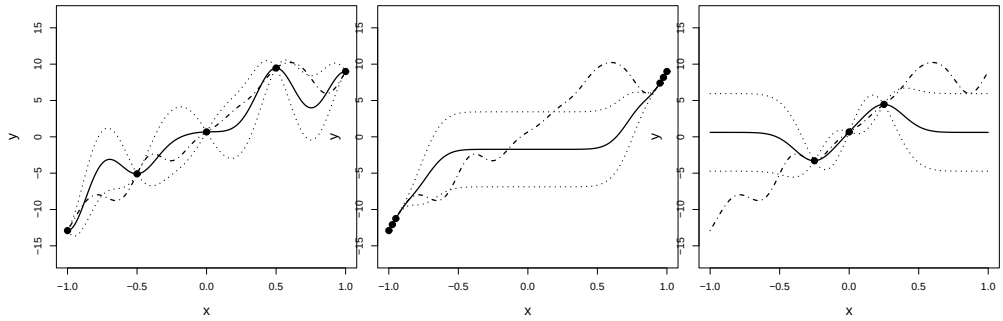

Figure 3: One realization (dashdot line) of a GP with linear trend is interpolated by Ordinary Kriging, based on 3 designs. The OK mean and 95% confidence intervals are represented by bold lines and dotted lines, respectively. The first design (left) is a regular grid; the associated OK prediction seems satisfying, even if the trend model is misspecified. The second design (center) is formed by 6 points concentrated at the boundaries of the domain; The Kriging predictor fails to capture the shape of the realization at the center of the domain. The third design is made of three points clustered at the center of the domain; OK automatically comes back to the mean value outside of the design and dramatically miss the actual trend.

We choose at first a subset of 5 regularly distributed points. Due to the fact that the points are regularly spaced on the grid, all the three kriging give similar good results, even if in two of the three cases the trend is misspecified (fig.3 left for the case of Ordinary Kriging). This may lead to the conclusion that specifying the trend is not very important and we could obtain good results using OK. But if we perform the same Krigings on different designs, where there are few points concentrated either on the boundaries or in the

center of the domain, then the results are very bad (due to the ratio between the parameter ψ_2 and the subdivision length) when the trend is misspecified, see (fig. 3, middle and right). The covariance parameters used for the simulated process in (fig. 3) and (fig. 4) are $\psi = (5, 0.2)$. The results are even

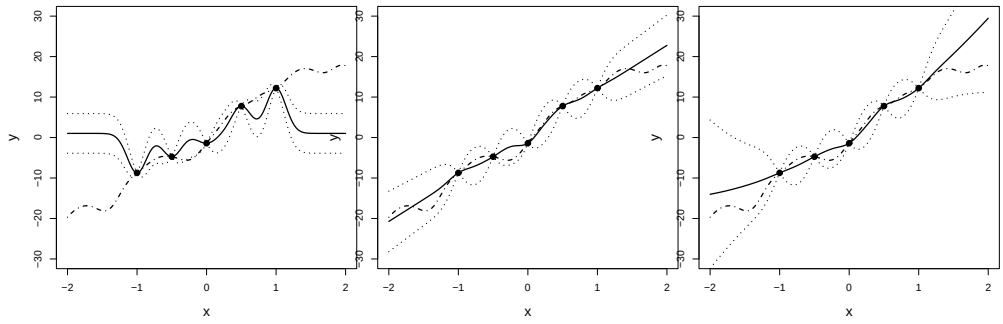

Figure 4: One realization (dashdot line) of a GP with linear trend is interpolated by three different Krigings, based on a regular grid (5 points evenly spaced between -1 and 1). The GP is here represented between -2 and 2, such that all three cases can be referred to as extrapolations. UK with affine trend (center) gives accurate results on the whole domain. On the contrary, both OK (left) and UK with quadratic trend (right) give good results between -1 and 1 but dramatically fail in extrapolation.

worse if we use the Kriging predictor given by OK or by UK with quadratic trend in extrapolation (fig.4, left and right). In the one dimensional case the choice of the trend doesn't seem to be essential while interpolating data which are not very distant one from another with respect to the frequency of variation of the process. On the contrary, when the design is not regular and we are in extrapolation, the performances of Kriging are very sensitive to the adequacy between the real trend of the process and the Kriging trend.

Hence it seems enough to properly fill the space to avoid the risks caused by the choice of trend functions. But what is possible in one or two dimensions becomes unrealistic when the dimension increases: a design with only one point at each vertex of a cubic domain $[0, 1]^d$ has 2^d points, i.e. 1024 points in 10 dimensions and more than a billion points in 30 dimensions. As we usually dispose of $10 \times d$ observations per dimension, which is already an optimistic case, choosing a trend based on data only appears as a very difficult task. Let us see nevertheless what would be possible in order to choose a trend starting from a data set $(\mathbf{X}, \mathbf{Z})$: the classical frame of linear regression offers a panel of diagnostic tools dedicated to validating both assumptions

on the trends and on the model of residuals. For instance, commonly used indicators include R^2 (and R^2 adjusted), the F-ratio and the p-values for each estimated regression coefficients, and numerous criteria to check the adequacy of the residuals to the underlying model. In most cases the Gaussian likelihood of the residuals is considered among the relevant criteria of model selection (some model testing techniques or even based upon it).

Now it seems necessary to recall that the latter measures are exclusively done at the design of experiments, also called "training sample" or "learning sets" in the literature of statistical learning, see [6]. Selecting only on the basis of a R^2 fit would lead for instance to the systematic choice of models interpolating ($\mathbf{X}$, $\mathbf{Z}$). However such models are not meant to be good in prediction outside the design of experiments. This warning leads to the double message:

- Model complexity must be taken into account in selection procedures
- Testing the model at some test points not used in the model fitting could be worth: this is for instance what cross-validation does.

The following experiment is performed in an intent to illustrate the first point. The second point will be illustrated in the next section. Here we investigate on a simple case how trend selection may be misleading when likelihood is the only criterion, without any consideration of model complexity. To do so, we compute, for each trend form of the Kriging model, the optimal parameter $\hat{p}$ by ML, we compare the corresponding values of the likelihoods and we select the kriging model having the highest value of likelihood. In table 3, we compare three Kriging models (OK, UK with linear trend, UK with quadratic trend) for three different functions: one realization of a one-dimensional GP with 11 points and with Gaussian covariance function ($\psi = (5, 0.4)$), the same realization plus a linear trend $0.5 + 5x$, and the same realization plus a quadratic trend $0.5 + 5x + 5x^2$.
Here it is essential to notice that the likelihood values are necessarily larger when adding more degrees of freedom to a statistical model. This constitutes a misleading incentive to always choose the model with the largest number of parameters within a given family. This happens for instance between Kriging models with first order and second order polynomial trends. As can be observed in table 3, L always increases (i.e. the values of $-2\ln(L(\hat{p}))$ will decrease) with the complexity of nested model. What we should really compare are maximum likelihood values between models with the same number

Table 3: Comparison of minimum $-2\log(L)$ values obtained by fitting three different Kriging models (OK, UK affine, UK quadratic) to three GP realizations. Each realization is drawn from the GP underlying one of the Kriging models. The design is a regular grid on $[-1, 1]$. The results illustrate that adding degrees of freedom to a Kriging model always lead to a larger value of the maximum likelihood.

	GP		GP +linear $\mathbf{t}$		GP+quadratic $\mathbf{t}$	
kriging type	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$
OK	0.4082	**32.07**	0.4445	36.90	0.4595	38.80
UK, linear $\mathbf{t}$	0.4085	**31.89**	0.4085	31.89	0.4387	35.80
UK, quadratic $\mathbf{t}$	0.4084	31.89	0.4084	31.89	0.4084	31.89

of degrees of freedom. On the last line of table 3, in the cases of the GP without trend and of the GP with linear trend , the estimated values $\hat{\beta}$ are very close to but different from zero. Thus the model obtained by automatically selecting the Kriging with highest likelihood will perform badly in extrapolation because of the higher order terms of the polynomial. The same phenomenon applies in an even more pronounced way with a linear trend in the case of a centered GP (first column, second row).

As a conclusion to this section, we have pointed out that OK and UK may seem to deliver similar results when the design is dense [22], but modeling the trend matters in extrapolation situations [9]. Since working in high-dimensional spaces means that we will practically always be in extrapolation, we need exploratory and visualization tools dedicated at finding trends in multivariate data. Recent methods of data mining and functional analysis may help [6]. We propose now to use additive models within spatial interpolation.

4 Using non-linear additive models as external drift

Linear models are often used by practitioners of quantitative disciplines since they are simple to interprete and to assess. Additive models (AM) are an extension of linear models. A precise description of these models can be found for instance in the book [5]. The advantage of AM is to conserve the feature of non-interacting predictors, but they allow much more flexible inference for each univariate problem, using kernel smoothers for instance [24]. The generic expression for an additive model is the following:

$$\begin{cases} Z_i = z(\mathbf{x}) + \varepsilon_i \\ z(\mathbf{x}) = \alpha + \sum_{j=1}^{d} f_j(\mathbf{x}_j) \\ \text{The } \varepsilon_i \text{ are } n.i.i.d. \end{cases} \tag{9}$$

and the f_js are arbitrary univariate functions, one for each predictor but possibly not the same kind of function for each dimension. Hence additive models deal with additive functions observed in a Gaussian noise. $\mathbf{x}$ is in fact assumed here to be a control variable, and not a random variable as in [5]. This model may be used to approximate deterministic computer experiments, provided that the response surface can reasonably be decomposed in an additive way. Once the nature of the f_js is chosen, they can be estimated using a powerful iterative procedure called *backfitting algorithm*, see [6]. Backfitting means that f_1 is estimated on the basis of all data $(\mathbf{X}, \mathbf{Z})$, then f_2 is fitted to the residuals $\mathbf{Z} - f_1(\mathbf{X})$, and so on. Under mild assumptions, the backfitting algorithm converges and finds the unique solution of the additive decomposition of (eq.10). In this section, we propose a combination of additive model and Kriging that offers the great flexibility of *AMs* and yet interpolates the data. It seems very natural to combine both models by using the following decomposition:

$$\begin{cases} z(\mathbf{x}) = t(\mathbf{x}) + \varepsilon_{SK}(\mathbf{x}) \\ t(\mathbf{x}) = \alpha + \sum_{j=1}^{d} f_j(\mathbf{x}_j) \\ \varepsilon_{SK}(\mathbf{x}) \text{ is a GP realization like in (eq.1)} \end{cases} \tag{10}$$

This identity may at first seem similar to the equation of Universal Kriging. However, in this case the non-linear nature of the trend prevents one from solving the estimation globally. Indeed, a likelihood maximization would lead to an optimization problem in infinite dimension:

$$\max_{\boldsymbol{\psi}, (f_j)_{j \in [1,d]}} L(\boldsymbol{\psi}, \mathbf{t}; \mathbf{Z}) \tag{11}$$

To our knowledge such a problem is analytically intractable. On the other hand, the backfitting algorithm is not suited anymore if we take the Kriging part into account. Indeed, kriging the residuals after fitting a smoother in one dimension would lead to an interpolation and thus end the iterative procedure without fitting the additive parts in the other dimensions.

Kriging with external trend [3] seems to constitute a good alternative for solving both the problem of the "general" form of trend and the one of the circularity. Consequently, we consider now a two-step approach (see Alg.1): first, the additive trend $t(\mathbf{x})$ is estimated using the backfitting algorithm,

and then Simple Kriging is applied to $(\mathbf{Z} - \mathbf{t})$ with covariance parameters estimated on the basis of those residuals, by likelihood maximization or other.

Algorithm 1 A first two-step approach to fit a Kriging with additive trend

1: Estimate the trend t by backfitting
2: Estimate the covariance parameters and fit a SK model on the basis of the residuals at $\mathbf{X}$

Unfortunately, there are significant drawbacks in the latter procedure, mainly related to the uncontrolled trade-off between deterministic and stochastic parts. Hence, the whole uncertainty reduces here to the Kriging variance estimated on the residuals; there is indeed no global uncertainty on the trend unless we use only splines in the *AM*. This is likely to cause a large underestimation of the process variance associated with the model. Furthermore, these residuals may be not very well suited to estimate the Gaussian process part: the additive model is constructed to fit z accurately at the design -possibly leading to *overfitting*-, thus the residuals at $\mathbf{X}$ are likely to vary with a smaller magnitude than in prediction. Since we look for a model with reasonable generalization properties, it seems necessary to find an alternative way of estimating the covariance parameters.

We propose here a sequential estimation technique for combined Kriging models like (eq.10). It is based on the idea that when the trend is non-linear, the parameters of the GP model should be estimated on a validation set rather than on the set at which the trend is fitted.

Algorithm 2 An alternative two-step approach to fit a Kriging with additive trend

1: Consider two designs $\mathbf{X}_1$ and $\mathbf{X}_2$ ▷ possibly obtained by splitting $\mathbf{X}$
2: Estimate the trend t by backfitting, based on the data $(\mathbf{X}_1, z(\mathbf{X}_1))$
3: Estimate the SK covariance parameters on the basis of the residuals $\{t(\mathbf{x}) - z(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}_2}$
4: Fit the SK model on the basis of all residuals ▷ with parameters estimated at the previous step

5 A 3-dimensional application of Kriging with Additive Trend (KAT)

The previous approach is applied to a 3-dimensional example from an industrial test case. The data are obtained with a flow simulator and the numerical response z, standing for the outcome of interest, is studied as a function of three physical parameters characterizing the porous media and denoted by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3 \in [-1, 1]$. The response is simulated at 1331 locations corresponding to a 11-level full factorial design, denoted by "F" in the sequel. Our goal is to provide a surrogate of the simulator on the basis of a poor design of experiments. The metamodel should interpolate the data (to respect the determinism of the underlying simulation) and provide a prediction uncertainty that allows statistical-based exploration, for instance to solve optimization problems. Furthermore, it should take into account a prior knowledge inherited from a previous study: the phenomenon is almost additive in its parameters.

Our initial design, "$\mathbf{X}_1$", is a 20-elements Hammersley sequence. We first perform a graphical analysis (fig. 5) of the response at $\mathbf{X}_1$, discuss the hypothese of additivity, and propose several kinds of linear and additive trends to model the data. Algorithm 1 is tested with the design $\mathbf{X}_1$ (Table 4). Then a second design, "$\mathbf{X}_2$", is used for an intermediate validation of the covariance parameters of the model previously obtained. $\mathbf{X}_2$ is made of 14 points taken from a 40-elements D-optimal design (see fig. 6). Algorithm 2 is then performed by re-estimating the covariance parameters of the previous SK model on the basis of the residuals at $\mathbf{X}_2$. An original estimation method is proposed, which differs from the traditional MLE: the process variance σ^2 is fixed such that the standardized residuals have most of their values between -2 and 2 [7] and the range parameter p is chosen in order to minimize the ISE at the design $\mathbf{X}_2$ (fig. 7). The full factorial design F is finally used for a phase of model validation (fig. 8).

A graphical analysis of the coplots at $\mathbf{X}_1$ does not reject the prior belief of additivity. A first additive decomposition is then estimated using splines in all directions (referred to as "GAM splines" in the following). We observe that we might take a linear trend in the directions of $\mathbf{x}_1$ and $\mathbf{x}_3$, and a non-linear trend in $\mathbf{x}_2$ without loosing much accuracy (see Table 4 for a quantitative validation). Hence we choose to fit an additive model with mixed trends, called "GAM mixed" in the sequel.

Different Krigings with external trend are fitted to the observed data at the

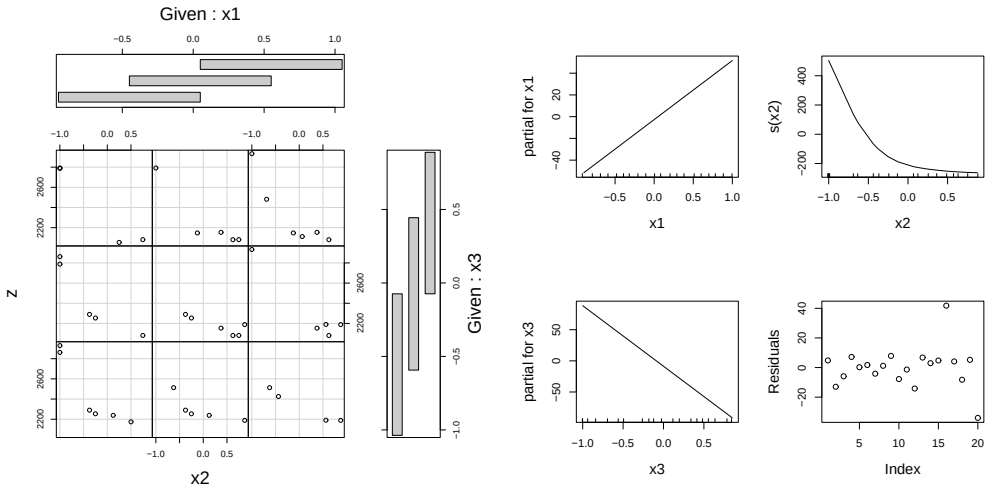

Figure 5: Coplots of z on the Hammersley design $\mathbf{X}_1$ (left) and summary of the additive components and the residuals obtained after application of the backfitting algorithm (right). The additive model is here chosen with a linear function in both directions of $\mathbf{x}_1$ and $\mathbf{x}_3$, and a smoothing spline in $\mathbf{x}_2$'s direction.

design $\mathbf{X}_1$. In all cases, the SK part has a structure of isotropic GP with Gaussian covariance (see section 2). We focus on the two additive trends defined above and on two additional linear trends: a first and a second order regression polynomial. For each model, we fit the trends respectively by OLS and backfitting, and we measure their relevance using indicators computed with the residuals at the design $\mathbf{X}_1$ (residuals deviance and p-values when available). Then we fit a Kriging to the residuals, as explained in Algorithm 1. For each Kriging, we store the maximum reached value of the log-likelihood and the corresponding range and variance values. The results are listed in (Table 4).

These results support the belief that a general additive trend is adapted for these data: both the variance of residuals and the values of their likelihood (compared to the 2nd order linear model, which uses more degrees of freedom) indicate their good fit to the data.

In practice, however, we care more about the model's abilities to make correct predictions at new points than about its mean squared error at the design. Hence, model validation should not be blindly supported by the indicator R^2 or the likelihood of the residuals at $\mathbf{X}_1$. First, we should consider the number of degrees of freedom of the model. Second, it may be worth validating the

Table 4: Optimal loglikelihood values and estimated covariance parameters associated with the residuals provided by Algorithm 1 at $\mathbf{X}_1$ with different trend structures. The R^2 values are computed by comparing the residual deviance after fitting the trend only, to the total variance of the response at the design $\mathbf{X}_1$.

Model	Loglikelihood	Range	σ^2	R^2_{adj}	R^2	p-value
1st order Linear + SK	-121.91	1.04	26101.89	0.78	0.82	4.03e-06
2st order Linear + SK	-100.69	0.048	1381.13	0.97	0.98	6.44e-08
GAM splines + SK	-76.01	0.048	117.10	-	0.99	-
GAM mixed +SK	-80.62	0.16	185.71	-	0.99	-

model outside of the design. Indeed, the residuals drawn from (fig. 5) are computed at the same locations as those used to fit the trend.

Concerning the first point, we can compare the degrees of freedom of both "2nd order linear" and "GAM mixed": respectively 10 and 7. Regarding the second point, we conduct a validation test on some additional data, inspired by the cross-validation procedure. Following Algorithm 2, $\mathbf{X}_2$ is used to valid and update the parameters associated with the model fitted at $\mathbf{X}_1$.

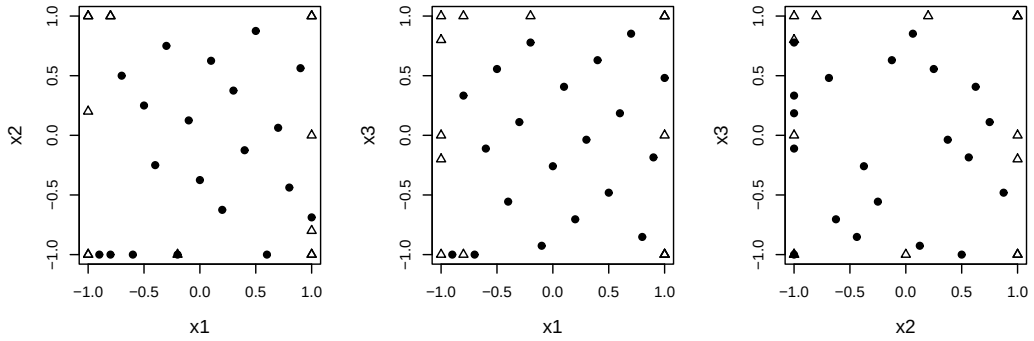

Figure 6: Coplots representing both $\mathbf{X}_1$ (dots) and $\mathbf{X}_2$ (triangles) designs in projection on all pairs of coordinates. The three graphics illustrate the space-filling behaviour of $\mathbf{X}_1$ and the D-optimal nature of $\mathbf{X}_2$. $\mathbf{X}_2$ also appears to be reasonably disconnected from $\mathbf{X}_1$.

The points of $\mathbf{X}_2$ are used to test the validity of the covariance parameters of the model "GAM mixed", previously estimated by ML. Figure 7 shows the associated residuals standardized by the ML variance (left), and the behaviour of the ISE at $\mathbf{X}_2$ as a function of p (right). We recall that the

residuals should satisfy the assumption of normality in order to get relevant Kriging variances, insuring correct statistical predictions.

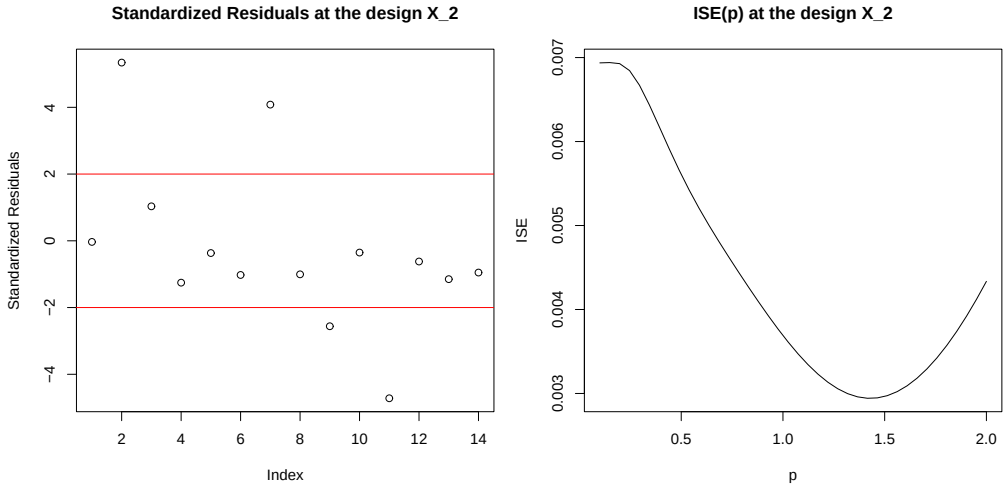

Figure 7: Standardized residuals at the validation design $\mathbf{X}_2$ (left) and variation of the ISE with respect to the covariance parameter p (right). One can observe that the value of p found by ML at $\mathbf{X}_1$ ($p_1 = 0.16$) is clearly suboptimal to get an accurate Kriging predictor when extrapolating to $\mathbf{X}_2$.

Figure 7 (right) shows that the ISE at the validation design $\mathbf{X}_2$ can be significantly reduced by increasing the range p. Following Algorithm 2, we re-estimate the covariance parameters based on these residuals at $\mathbf{X}_2$. Instead of using ML however , we prefer to directly use the work done hereabove to compute the ISE as a function of the range. It appears indeed that the optimal range to accurately fit the residuals at the validation design is given by $p_2 = 1.4$. Concerning the variance, we observe more satisfying standardized residuals with $\sigma_2^2 = (2 \times \sigma_{ML})^2$. So we keep σ_2^2.

Remark: A ML estimation with the residuals at $\mathbf{X}_2$ delivers $p = 0.97$.

We finally test the model of Algorithm 2 at the design F (fig. 8). The standardized residuals (with the variance σ_2^2) and the ISE as a function of p validate our empirical decisions made on the basis of the intermediate design $\mathbf{X}_2$ (note that ML on $\mathbf{X}_2$ -see remark hereabove- gives also better results than ML on $\mathbf{X}_1$ but the cross-validating strategy minimizing the ISE at $\mathbf{X}_2$ remained the best). To conclude with, the algorithm investigated performed well on this example: Simple Kriging seems to constitute a good complement to additive models in an intent to interpolate data and also possibly

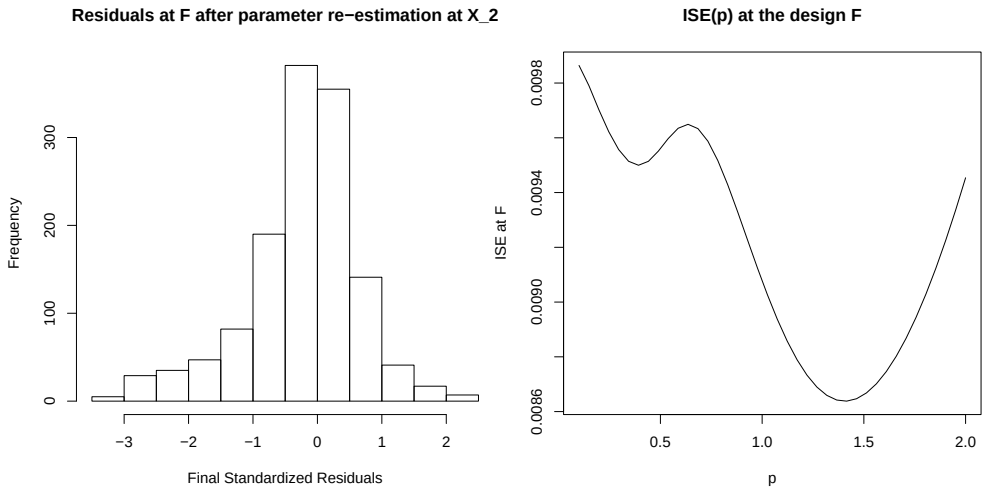

Figure 8: **Left:** Histogram of the standardized residuals at the test design F with the model previously obtained by Algorithm 2 (GAM mixed, $\sigma^2 = \sigma_2^2,\ p = p_2$). **Right:** Variation of the ISE with respect to the covariance parameter p: the value $p_2 = 1.4$ previously chosen at $\mathbf{X}_2$ is almost optimal again.

explain some non-additive part. The method we use here allows inference of covariance parameters with values suited for a correct quantification of uncertainty. This seems encouraging to develop further "cross-validation-like" methods for the combination *Additive model + Kriging*.

6 Conclusions and perspectives

We observed in a one-dimensional frame that MLE could behave very differently from Fisher asymptotical results when n is small. This result should be kept in mind when dealing with higher dimensions, and further studies have to be done in this latter context. Since it relies on the simulation of Gaussian vectors, the experimental approach presented here can easily be transposed in a higher dimensional framework. Perspectives include the empirical comparison of ML and penalized ML [11] when using classical designs.

Further experiments on the topic of trend selection illustrated the fact that the likelihood cannot be considered as only criterion when comparing different functional families. This is suggesting methods penalizing complexity (like in AIC and BIC). But we mainly wish to emphazise on the risks took

when predicting with trended Kriging: in higher dimensions, we will allways be in an extrapolation situation. Choosing a trend with the help of a small design then seems very risky. This is an argument to consider Ordinary Kriging in the cases where no prior information on the trend is available.

In other respects, we proposed a model combining an additive model and Simple Kriging. The application to a simple industrial test case confirmed that directly kriging the residuals by ML gives a poor result. Our attempt to adapt a method inspired by cross-validation with a single test set gave here a Kriging with different features from ML, apparently accounting well for the non-additive part of the response. However, the question of the robustness to a change of design has not been raised yet. This is a subject to be treated in further works.

Aknowledgements

All the computations have been performed using *R* [17] and the packages *geoR* [18], *RandomFields* [21], *gam* and *gstat*. This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA, and Total S.A.

References

[1] Markus Abt. Estimating the prediction mean squared error in gaussian stochastic processes with exponential covariance structure. *Scandinavian Journal of Statistics*, 26:563–578, 1999.

[2] Markus Abt and William J. Welch. Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes. *The Canadian Journal of Statistics*, 26:127–137, 1998.

[3] N.A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics, 1993.

[4] Sarah Goria. *Evaluation d'un projet minier: approche bayésienne et options réelles*. PhD thesis, Ecole des Mines de Paris, 2004.

[5] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1991.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[7] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[8] Astrid Jourdan. Approches statistiques des expériences simulées. *Revue de Statistiques Appliquées*, 50:49–64, 2002.

[9] A. G. Journel and M. E. Rossi. When do we need a trend model in kriging? *Mathematical Geology*, 21(7):715–739, 1989.

[10] J.R. Koehler and A.B. Owen. Computer experiments. Technical report, Department of Statistics, Stanford University, 1996.

[11] R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, (47):111–120, 2005.

[12] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–46, 1984.

[13] J.D. Martin and T. W. Simpson. A monte carlo simulation of the kriging model. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY*, page 4483, 2004.

[14] J.D. Martin and T.W. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43 (4):853–863, 2005.

[15] Georges Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.

[16] A. O'Hagan. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, (91):1290–1300, 2006.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[18] P.J. Ribeiro Jr. and P.J. Diggle. geoR: A package for geostatistical analysis, 2001. ISSN 1609-3631.

[19] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, (4):409–435, 1989.

[20] T.J. Santner, B.J. Williams, and W.J. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[21] M. Schlather. Simulation and analysis of random fields, 2001. URL: http://www2.hsu-hh.de/schlath/R/RandomFields.

[22] Michael L. Stein. *Interpolation of Spatial Data, Some Theory for Kriging*. Springer, 1999.

[23] T.J. Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, 8 (6):1375–1381, 1980.

[24] Grace Wahba. *Spline Models for Observational Data*. S.I.A.M., 1990.